Gender-Based Differential Prediction by Curriculum Samples for College Admissions

Niessen, A. Susan M.; Meijer, Rob R.; Tendeiro, Jorge N.

Published in:
Educational Measurement: Issues and Practice

DOI:
10.1111/emip.12266

Publication date:
2019

Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2019, Vol. 38, No. 3, pp. 33–45

Gender-Based Differential Prediction by Curriculum Samples for College Admissions

A. Susan M. Niessen, Rob R. Meijer, and Jorge N. Tendeiro, *University of Groningen*

A longstanding concern about admissions to higher education is the underprediction of female academic performance by admission test scores. One explanation for these findings is selection system bias, that is, not all relevant KSAOs that are related to academic performance and gender are included in the prediction model. One solution to this problem is to include these omitted KSAOs in the prediction model, many of these KSAOs are 'noncognitive' and "hard-to-measure" skills in a high-stakes context. An alternative approach to capture relevant KSAOs is using representative performance samples. We examined differential prediction of first year- and third year academic performance by gender based on a curriculum-sampling test that was designed as a small-scale simulation of later college performance. In addition, we examined differential prediction using both frequentist and Bayesian analyses. Our results showed no differential prediction or small female underprediction when using the curriculum-sampling tests to predict first year GPA, and no differential prediction for predicting third year GPA. In addition, our results suggest that more comprehensive curriculum samples may show less differential prediction. We conclude that curriculum sampling may offer a practically feasible method that yields minimal differential prediction by gender in high-stakes operational selection settings.

Keywords: Bayesian analyses, college admission, curriculum sampling, differential prediction, high-stakes assessment

Having a college degree determines to a large extent an individual's employment opportunities and is "more indicative of income, of attitudes, and of political behavior than . . . region, race, age, religion, sex and class" (Lemann, 1999, p. 6). It is thus extremely important for individuals and society that admission procedures to higher education are fair and are not biased against, for example, gender, ethnicity, or socioeconomic status. In this context, we differentiate between adverse impact and bias. Adverse impact is defined as systematic differences in scores, and thus chances of acceptance between subgroups. Bias is defined as differential prediction; a procedure is biased when there are systematic differences in criterion performance between subgroups, conditional on the admission test score (Guion, 1998). Adverse impact can be a sign of bias, but this is not necessarily the case (e.g., Kuncel & Hezlett, 2010). The focus of this study is differential prediction. Differential prediction is usually studied using moderated multiple regression models, examining differences in intercepts and slopes between prediction models for different subgroups (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 66; Cleary, 1968). Conclusions about the absence of differential prediction are often drawn based on these analyses. However, such conclusions are problematic based on frequentist regression analyses, which are typically also underpowered (Aguinis, Culpepper, & Pierce, 2010). Therefore, we also included Bayesian-step-down regression analyses in this study, which allowed us to quantify the evidence in favor of and against the occurrence of differential prediction (Kruschke, Aguinis, & Joo, 2012).

Differential Prediction in Admission Testing

Given the major interests that are at stake, it is not surprising that differential prediction is a well-researched area in preemployment testing and in college admission testing. An often-reported finding is that admission test scores show underprediction for female applicants and overprediction for ethnic minority applicants; that is, female applicants obtain better academic results than predicted by their admission test scores, whereas ethnic minority applicants obtain lower academic results than predicted by their admission test scores (e.g., Fischer, Schult, & Hell, 2013; Keiser, Sackett, Kuncel, & Brothen, 2016; Mattern & Patterson, 2013; Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008; Schult, Hell, Päßler,

A. Susan M. Niessen is an Assistant Professor at the Department of Psychology, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1 9712TS Groningen, The Netherlands; a.s.m.niessen@rug.nl. Rob R. Meijer is a Full Professor at the Department of Psychology, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1 9712TS Groningen, The Netherlands; r.r.meijer@rug.nl and Jorge N. Tendeiro is an Assistant Professor at the Department of Psychology, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1 9712TS Groningen, The Netherlands; j.n.tendeiro@rug.nl

& Schuler, 2013; Shewach, Shen, Sackett, & Kuncel, 2017). Several explanations for these differential prediction findings have been proposed. However, it is still unclear which mechanisms underlie differential prediction by ethnicity (Aguinis, Culpepper, & Pierce, 2016; Mattern et al., 2008) but language seems to be a factor (Shewach et al., 2017).

Female Underprediction

With respect to female underprediction Fischer et al. (2013) showed that the amount of underprediction was unrelated to the magnitude of predictor or criterion score differences. Thus bias in the test or the criterion is likely not the cause of female underprediction (Meade & Fetzer, 2009). Therefore, Fischer et al. (2013) and several other authors argued that female underprediction is not caused by bias in admission tests themselves but by "selection system bias," that is, the omission of valid variables from the prediction model that are related to the criterion variable (e.g., college GPA) and to gender (Fischer et al., 2013; Jencks, 1998; Sackett, Laczo, & Lippe, 2003). In other words, admission procedures do not include all knowledge, skills, abilities, and other factors (KSAOs) that are (a) relevant for college performance and (b) related to gender.

Most proposed omitted KSAOs are "noncognitive,"[1] such as motivation and study habits (e.g., Fischer et al., 2013). Many studies have shown that noncognitive traits and skills, such as conscientiousness, academic discipline, motivation, and study skills and study habits are related to academic performance in higher education (e.g., Borghans, Golsteyn, Heckman, & Humphries, 2016; Busato, Prins, Elshout, & Hamaker, 2000; Credé & Kuncel, 2008; Richardson, Abraham, & Bond, 2012), and to gender, with higher scores for females (De Bolle et al., 2015; Duckworth & Seligman, 2006; Schmitt, Realo, Voracek, & Allik, 2008; Strenta, Elliot, Adair, Matier, & Scott, 1994). Indeed, Stricker, Rock, and Burton (1993) showed that female underprediction was reduced when they included self-reported "studious behavior" in the prediction model. More recently, Keiser et al. (2016) and Kling, Noftle, and Robins (2012) found that adding conscientiousness scores to prediction models containing standardized admission test scores reduced female underprediction. Similarly, Mattern, Sanchez, & Ndum (2017) found that female underprediction was reduced when adding a measure of academic discipline to a model containing ACT scores and high school GPA. Consequently, several researchers (e.g., Goldstein, Zedeck, & Goldstein, 2002; Keiser et al., 2016; Mattern et al., 2017) recommended the inclusion of such "omitted" KSAOs in admission procedures. However, these authors also acknowledged that assessing such "noncognitive" KSAOs in high-stakes admission procedures is challenging. Most existing assessments rely on self-reports, and there are almost no studies that investigated the effectiveness of noncognitive predictors such as personality traits or study skills and habits in high-stakes procedures. Due to faking, the generalizability of the validity of scores obtained in low-stakes conditions on such instruments to high-stakes admission conditions is limited (Niessen, Meijer, & Tendeiro, 2017b; Peterson, Griffith, Isaacson, O'Connell, & Mangos, 2011).

Approaches to Assessing Relevant KSAOs

The traditional approach to assessing different KSAOs is to define and measure them as distinct constructs; the studies cited above provide good examples of this approach. Wernimont and Campbell (1968) refer to this approach as a "signs" approach to prediction. An alternative is to use representative samples of relevant performance, based on the idea of behavioral consistency. Wernimont and Campbell (1968) called this a "samples" approach to prediction. The idea is that when such performance samples are representative enough, they should tap into the same cognitive and noncognitive KSAOs that are relevant for the criterion performance (e.g., Callinan & Robertson, 2000; Hough, Oswald, & Ployhart, 2001; Lievens & De Soete, 2012). So, performance samples are multifaceted performance measures; relevant KSAOs are not measured in isolation, but within representative tasks that often require a mixture of cognitive and noncognitive skills (Callinan & Robertson, 2000; Hough et al., 2001). If this is indeed the case, performance samples may be a very suitable way to deal with the omitted variables problem that leads to selection system bias and differential prediction.

Reducing differential prediction by using a samples approach has been suggested previously in the personnel selection literature (e.g., Aramburu-Zabala Higuera, 2001; Ployhart & Holtz, 2008; Robertson & Kandola, 1982). In the context of higher education, the lower differential prediction of high school GPA as compared to other admission criteria (Fischer et al., 2013; Mattern et al., 2008; Zwick, 2017; Zwick & Himmelfarb, 2011) may also be explained based on this rationale. While not a "sample" of college performance, high school GPA is also a multifaceted performance measure that taps into cognitive skills and abilities, knowledge, and the ability to "get it done" (Bowen, Chingos, & McPherson, 2011, p. 123; Borghans et al., 2016; Deary, Strand, Smith, & Fernandes, 2007). However, there are many practical drawbacks to using high school GPA in admission procedures, such as negative applicant reactions (Niessen, Meijer, & Tendeiro, 2017a), and different grading practices across high schools and countries, leading to comparability problems (Zwick, 2017, p. 57).

A Samples Approach to Admission Testing

In Europe,[2] applicants are increasingly selected on the basis of curriculum-sampling tests (de Visser et al., 2017; Häkkinen, 2004; Lievens & Coetsier, 2002; Niessen, Meijer, & Tendeiro, 2016, 2018; Reibnegger et al., 2010; Vihavainen, Luukkainen, & Kurhila, 2013). These curriculum-sampling tests are designed based on the same rationale as work-sample tests used in personnel selection (e.g., Callinan & Robertson, 2000). For example, in the Netherlands the curriculum-sampling approach was first implemented after some studies found that the grade on the first course in the program, usually an *Introduction to . . .* course, was a very good predictor of later academic performance (e.g., Busato et al., 2000; Korthals, 2007; Niessen et al., 2016). The idea was to design an admission test that served as a small-scale simulation of such an *Introduction to . . .* course, because such a test was expected to have good predictive validity, and could also offer the applicants some insight into the content of the program.

For admission to most undergraduate programs, a curriculum-sampling test usually consists of studying college-level domain-specific material and taking an exam, mimicking what is often required in undergraduate programs at research universities. A main difference with admission tests such as the SAT or ACT is that the test is matched to the program of interest in content and form (e.g., Sackett, Walmsley, Koch, Beatty, & Kuncel, 2016). A difference with tests such as SAT

subject tests and AP exams is that curriculum samples are not designed to assess prior knowledge or skills obtained at the high-school level, but require applicants to study college-level material that they are not yet familiar with. Thus, the material and the exam that they encounter are on the first-year college level, and require similar preparatory activities as an exam within the program. In that sense, curriculum samples are simulations of the college program. Hence, curriculum samples can be used in admission procedures to specific programs, such as undergraduate and graduate programs in Europe and graduate programs and specialized majors in the United States. In addition, we note that the curriculum-sampling approach can also be used to assess practical skills for practice-oriented programs, such a computer science or teacher education (for some examples, see Valli & Johnson, 2013; Vihavainen et al., 2013).

Previous studies found that curriculum-sampling tests were good predictors of academic performance with uncorrected correlations with 1YGPA ranging between .40 and .50 (Niessen et al., 2016, 2018), and better performance and lower dropout rates for applicants admitted through this method (Booij & van Klaveren, 2017; de Visser et al., 2017; Reibnegger et al., 2010; Visser, van der Maas, Engels-Freeke, & Vorst, 2012). In addition, Niessen et al. (2017a) found that applicants perceived curriculum samples more favorably than many other admission methods. The high similarity to the criterion performance, and the measurement of the same cognitive- and noncognitive KSAOs needed for good criterion performance, has often been suggested as an explanation for the favorable validity of sample-based assessments in personnel selection (Asher & Sciarrino, 1974; Callinan & Robertson, 2000; Lievens & De Soete, 2012). In the context of higher education, Lievens and Coetsier (2002) found that curriculum-sampling test scores were related to cognitive abilities, and to a smaller extent to personality traits. In contrast, Niessen et al. (2018) found no relationships between scores on curriculum-sampling tests and scores on a cognitive ability test. However, curriculum-sampling test scores were related to some noncognitive constructs (e.g., conscientiousness, time management) that were also related to academic performance in a psychology program. Following this rationale, using representative performance samples may also lead to little or no differential prediction, due to tapping into relevant KSAOs. However, to our knowledge, there are no studies that investigated differential prediction of performance or curriculum samples.

Aim of the Present Study

In the present study, we investigated differential prediction by gender using curriculum samples as predictors of academic performance. We hypothesized that representative curriculum-sampling tests should show no or trivial differential prediction, because they tap into the same KSAOs that are associated with successful performance in the college program. In addition, more comprehensive performance samples tend to have higher predictive validity, because they tap into relevant KSAOs more effectively (e.g., Callinan & Robertson, 2000). Accordingly, we hypothesized that more comprehensive curriculum samples would show less differential prediction. To investigate this hypothesis, we studied differential prediction by gender for a curriculum-sampling test used for admission to a psychology program. The test was designed as a small-scale version of the *Introduction to Psychology* course. We also investigated differential prediction by gender for the grade of the *Introduction to Psychology* course, which can be viewed as a more comprehensive sample of the curriculum. This latter predictor is not practically feasible as an admission test, but served to explore our hypothesis that differential prediction would be reduced when the representativeness and comprehensiveness of the curriculum sample increased. We investigated differential prediction in three cohorts using first year GPA as the criterion measure. In addition, we did the same analyses using third year GPA as the criterion variable, which was available for one cohort.

To study these expectations we used both a frequentist and a Bayesian (e.g., Kruschke et al., 2012) step-down regression approach (Lautenschlager & Mendoza, 1986). A step-down moderated multiple regression consists of three steps. First, an omnibus test for slope and intercept differences is performed. If the results indicate differential prediction, a test for slope differences is performed as a second step and a test for intercept differences is performed as a third step. The procedure is described in more detail in the Method section.

A Bayesian approach is particularly suitable in this study because, contrary to the frequentist approach, it allows us to examine the evidence in favor of the null hypothesis of no differential prediction. For example, contrary to the interpretations in some studies (e.g., Hough et al., 2001), the absence of statistically significant slope differences based on frequentist analyses does not imply that they are nonexistent, especially given the low power for detecting slope differences in most studies (e.g., Aguinis et al., 2010). Using a Bayesian approach, we can quantify how much the data support the null hypothesis of no slope differences. So, the aim of this paper was twofold: First, we investigated if a curriculum-sampling approach would indeed show no or minimal differential prediction by gender in a high-stakes context. Second, we used a Bayesian approach to differential prediction analyses to illustrate how this technique can contribute to the interpretation of differential prediction results and thus to the sound development of differential prediction analyses.

Method

Participants and Procedure

The samples included applicants to an undergraduate psychology program at a Dutch university. The data consisted of applicants who applied to the program in 2013, 2014, or 2015, and who subsequently enrolled in the program and participated in at least one course. All participants completed a curriculum-sampling test in the admission procedure. The admission procedure also consisted of an English-reading comprehension test and a math test in 2013 and 2014, and of a math test and a test about material provided through a video lecture in 2015. The admission committee did not reject any applicants, because the number of applicants who did not withdraw their application did not exceed the number of available places. However, this was not known beforehand and the procedure was thus perceived as high-stakes.[3] The students followed the study program either in English or in Dutch, with similar content. The majority of the students who followed the English program were international students, mostly from Germany. Some international applicants were allowed to take the admission tests online (13%, 16%, and 16% of all test-takers, respectively). Since test administration was

not proctored in these cases, we removed these cases from the data set. All data were obtained through the university administration. This study was approved by and in accordance with the rules of the Ethical Committee Psychology from the university.

Sample 1. The first sample consisted of the 576 applicants who applied to the program and enrolled in 2013. Sixty-nine percent was female and the mean age was $M = 20$ ($SD = 2.1$). The Dutch program was followed by 45% of the students. The nationalities of the applicants were 49% Dutch, 41% German, 8% other European countries, and 2% non-European.

Sample 2. The second sample consisted of the 552 applicants who applied to the program and enrolled in 2014. Sixty-five percent was female and the mean age was $M = 20$ ($SD = 1.6$). The Dutch program was followed by 45% of the students. The nationalities were 47% Dutch, 45% German, 7% other European countries, and 1% non-European.

Sample 3. The third sample consisted of the 471 applicants who applied to the program and enrolled in 2015. Seventy percent was female and the mean age was $M = 20$ ($SD = 2.0$). The Dutch program was followed by 42% of the students. The nationalities were 47% Dutch, 44% German, 8% other European countries, and 1% non-European.

Measures

Curriculum-sampling test. The curriculum-sampling test was designed to mimic the first course in the program: *Introduction to Psychology.* The applicants had to study two chapters of the book used in this course, which they could access since January. Hence, the time the students had to prepare for the test was not restricted. On the selection day, which took place at the university in May or June, they took a multiple-choice exam about the material, because this is the most common type of exam in this program. The applicants had 45 minutes to complete the exam, which was constructed by a course instructor. Each year, the exams consisted of different items (40 items in 2013 and 2014, 39 items in 2015); the estimated reliability of the tests was $\alpha = .81$ in 2013, $\alpha = .82$ in 2014, and $\alpha = .76$ in 2015.

Introduction course grade. The grade in the course *Introduction to Psychology* obtained in the program qualifies as the result of a more comprehensive curriculum sample than the admission test. This was the grade obtained at the first exam of the course (not including resit scores). The course covered similar, but more comprehensive content than the curriculum-sampling test. During the first half of the first semester, students attended nonmandatory lectures, studied a book, and took a multiple-choice exam about the material. The reliability of the exam was $\alpha = .74$ in 2013, $\alpha = .81$ in 2014, and $\alpha = .77$ in 2015. The exam was graded on a scale ranging from 1 to 10. In each cohort, some students did not participate in this exam, leading to missing values (2%, 1%, and 2%, respectively). The missing values were handled by listwise deletion in all analyses.[4] The exact sample sizes for all variables in each sample are shown in Table 1 (column 5).

First year GPA. First year GPA (1YGPA) was the mean grade obtained after one academic year; there were 10 course grades when a student completed all courses. Grades were given on a scale from 1 to 10, with a 6 or higher representing a pass. For most courses, literature had to be studied on psychological or methodological topics, supplemented with noncompulsory lectures, and assessed through a multiple-choice exam. For analyses including both the introduction course grade and 1YGPA, the grade on the first course was excluded from 1YGPA to avoid inflation of the validity coefficients.

Third year GPA. Third year GPA (3YGPA) was available for 450 participants from the first sample; the other students dropped out of the program. 3YGPA was defined as the mean grade obtained after three academic years. The number of courses completed by each student varied, but students were expected to complete the undergraduate program within three years. The courses in the first and second year were mostly the same for all students, whereas the third year consisted of mostly elective courses in subdisciplines of psychology.

Frequentist and Bayesian Approach

There were several reasons to supplement the classical frequentist analyses with a Bayesian approach (e.g., Gelman et al., 2014; Kruschke et al., 2012). First, there are some shortcomings of the classical step-down regression analysis (Lautenschlager & Mendoza, 1986) to study differential prediction (Aguinis et al., 2010; Berry, 2015; Meade & Fetzer, 2009). Tests for slope differences tend to be underpowered, even in large samples, and tests for intercept differences tend to have inflated Type I errors (Aguinis et al., 2010). There have been suggestions to overcome these problems (Aguinis et al., 2010; Berry, 2015; Mattern & Patterson, 2013; Meade & Fetzer, 2009), but most suggestions rely on visual inspection, or assume that there are no slope differences, or that they are difficult to implement (for example, improving test reliability and reducing subgroup score differences; e.g., Berry, 2015). A Bayesian approach does not solve all these problems, but inconclusive results can be distinguished from evidence in favor of the null hypothesis of no differential prediction. Second, the Bayesian approach provides comprehensive tools for parameter estimation and hypothesis testing (e.g., Gelman et al., 2014; Kruschke et al., 2012). Through Bayesian statistics probabilities for model parameters can be computed after observing the data, thus $p(\text{theory}|\text{data})$ can be computed. Under the classical frequentist framework, a researcher usually computes the probability of observing the data at hand or more extreme given that the model under consideration holds, that is, $p(\text{data}|\text{theory})$. Most researchers are, however, interested in assessing the plausibility of research hypotheses based on the observed. In that case, Bayesian statistics typically provide direct answers. Under the frequentist approach, we cannot compute $p(\text{theory}|\text{data})$ because theories have no stochastic properties, only data do. A third reason for using a Bayesian approach is that it does not capitalize on issues such as dependence on unobserved data, subjective stopping data collection rules, multiple testing, and lack of support for the null hypothesis (Gelman et al., 2014; Wagenmakers, 2007).

In the present study, the Bayesian approach thus has the advantage that when we use different types of curriculum samples as predictors we can investigate whether the data are more in agreement with the hypothesis that no differential prediction occurs (i.e., the null hypothesis). In addition, contrary to well-known confidence intervals (CIs) used in the frequentist approach, credible intervals based on Bayesian analyses (BCIs) can be interpreted as the most probable values of a parameter given the data (e.g., Kruschke

Table 1. Means, Standard Deviations, and Gender differences

Variable	Sample	Overall			Men		Women		σ	BF_{10}
		M	*SD*	*n*	*M*	*SD*	*M*	*SD*		
Curriculum sample	2013	29.80	5.0	576	29.40	5.44	29.98	4.89	–.11 [–.29, .07]	.22
	2014	29.94	5.45	552	29.86	5.62	29.99	5.37	–.02 [–.20, .15]	.10
	2015	29.25	4.61	471	30.03	3.92	28.92	4.84	.23 [.04, .42]	1.79
Introduction course grade	2013	6.67	1.36	565	6.44	.51	6.77	1.27	–.24 [–.41, –.06]	3.29
	2014	6.74	1.65	547	6.73	1.75	6.74	1.59	–.01 [–.18, .16]	.10
	2015	6.30	1.51	461	6.37	1.64	6.27	1.45	.06 [–.13, .26]	.14
1YGPA	2013	6.63	1.29	576	6.32	1.46	6.78	1.18	–.35 [–.53, –.17]	215.48
	2014	6.44	1.34	552	6.38	1.34	6.48	1.34	–.07 [–.24, .10]	.14
	2015	6.64	1.23	471	6.59	1.31	6.66	1.20	–.05 [–.24, .14]	.13
3YGPA	2013	7.09	.72	450	7.04	.80	7.12	.69	–.10 [–.31, .11]	.19

Note: σ is the estimated effect size based on the Bayesian analysis, 95% credible intervals are between brackets. BF_{10} shows the Bayes factor for the evidence in favor of the alternative hypothesis relative to the null hypothesis. Men were coded 0, women were coded 1.

& Liddell, 2018). We, therefore, decided to use Bayesian techniques in our analyses and we compared the frequentist results with the Bayesian results.

Analyses

Means and standard deviations were inspected, and corresponding effect sizes for the differences between means for male and female applicants were calculated for all samples. For each predictor–criterion combination, we conducted step-down hierarchical regression analyses (Lautenschlager & Mendoza, 1986), which is a commonly used and recommended approach to differential prediction analysis (Aguinis et al., 2010). This procedure starts with an omnibus test that is used to compare a simple regression model that only includes the main continuous predictor (the curriculum-sampling test score or the introduction course grade) with a regression model, that includes the continuous predictor, gender, and a predictor–gender interaction term. If the result of the omnibus test is indicative of differential prediction (i.e., if the *p*-value is below the prestipulated 5% significance level for frequentist analyses, or the Bayes factor indicates evidence in favor of differential prediction for Bayesian analyses), subsequent sequential tests of slope differences and intercept differences are conducted. Slope differences are determined through testing a regression model including the first-order continuous predictor and gender effects against a full regression model also including an interaction term. When slope differences are detected, intercept differences are assessed through testing a regression model that includes the continuous predictor and the predictor–gender interaction term against a full model also including the gender main effect. To reduce multicollinearity, the independent variables were centered around their means before analyses were conducted (Cohen, Cohen, West, & Aiken, 2003). Because we examined both predictors with 1YGPA as the criterion measure in three samples and with 3YGPA as the criterion measure in one sample, eight step-down regression analyses were conducted with both approaches.

Frequentist analyses. For the frequentist analyses, an alpha level of .05 per test was chosen. In addition to regression coefficients and ΔR^2 values, we computed d_{Mod} standardized effect sizes for the degree of differential prediction as recommended by Dahlke and Sackett (2018). Typically, in differential prediction studies ΔR^2 or (standardized) differences in mean residuals between groups are used as effect sizes. However, standardized mean residuals may cancel each other out when slope differences are present, and are affected by differences in predictor scores between groups. Similarly, ΔR^2 does not show the direction of the effect (Nye & Sackett, 2017). Nye and Sackett (2017) proposed several d_{Mod} effect sizes that do not have these limitations and they showed the practical value of differential prediction effects in standardized metrics. As recommended, we computed d_{Mod_signed} (the signed effect size for differential prediction), $d_{Mod\_unsigned}$ (the unsigned effect size for differential prediction, that does not allow canceling out due to slope differences), d_{Mod_under} (the standardized difference in prediction in the score range where negative differences in prediction occurred), d_{Mod_over} (the standardized difference in prediction in the score range where positive differences in prediction occurred), $d_{Mod\_max.}$ (the largest absolute-value difference between groups' regression lines), and the proportion of over- and underpredicted female applicants in each cohort. d_{Mod} effect sizes can be interpreted similarly to Cohen's *d* (Dahlke & Sackett, 2018). They were computed using the *psychmeta* package in R (Dahlke & Wiernik, 2018).

Bayesian analyses. For the Bayesian analyses, the Bayes factor (Kass & Raftery, 1995) was used as a measure of evidence for or against differential prediction at each step in the regression analyses (Lautenschlager & Mendoza, 1986). The Bayes factor shows the weight of evidence in the data for competing hypotheses, or the degree to which one hypothesis predicts the observed data better than the other. For example, a Bayes factor of H_1 against H_0 of 3 (denoted $BF_{10} = 3$) means that the empirical data is three times more likely to occur under H_1 than under H_0; $BF_{10} = 1$ means that the empirical data are equally likely under both hypotheses (e.g., Gelman et al., 2014; Kass & Raftery, 1995).

To interpret the Bayes factors we used the benchmarks proposed by Kass and Raftery (1995, p. 777).[5] The Bayesian analyses were conducted using the R package *BayesFactor* (Morey & Rouder, 2015) to compute the Bayes factors, and using JAGS, version 4.2.0 (Plummer, 2016a) in R, with the package rjags, version 4.6 (Plummer, 2016b) for model estimation.

Bayesian analysis starts by specifying a prior distribution for the parameters. After data collection, a posterior distribution combining information from the data and the prior is computed. Posterior distributions cannot be calculated directly, so the posterior distribution is approximated based on

Markov chain Monte Carlo (MCMC) sampling (for details, see Kruschke et al., 2012). The default priors used by function *regressionBF* in the *BayesFactor* R package were used. This is a Jeffreys prior on the joint distribution for the intercept and errors variance, and a particular normal prior for the regression coefficients (for details, see Liang, Paulo, Molina, Clyde, & Berger, 2008). For model estimation, we used broad priors: a normal prior on the standardized regression coefficients with a mean of zero and a standard deviation of 100, and a uniform prior on the residual variance ranging from zero to ten. The standardized regression coefficients were transformed back to the original scale. We used 1,000 iterations to tune the samplers and 1,000 burn-in iterations before running four MCMC chains of 10,000 iterations each. Convergence of the MCMC iterations (Gelman-Rubin's convergence diagnostic) and effective sample size were inspected and no problems were detected.

Results

Table 1 shows descriptive statistics for the curriculum-sampling test, the introduction course grade, 1YGPA, and 3YGPA in each cohort, for men and women, and effect sizes for the difference in scores between men and women. We only reported the results based on the Bayesian approach, because results differed very little when obtained with a frequentist approach. All differences in scores between men and women were small. When we inspect the Bayes factors, there was anecdotal evidence (Kass & Raftery, 1995) that men performed better than women on the curriculum-sampling test in 2015 ($BF_{10} = 1.73$), but all credible values for the effect size of the difference were small (95% BCI [.04, .42]). There was positive evidence that women performed better than men in the introduction course in 2013 ($BF_{10} = 3.29$) and strong evidence that women performed better than men in the first year in 2013 ($BF_{10} = 215.48$), both with small to moderate credible effect sizes (95% BCI [–.41, –.06] and [–.53, –.17], respectively). Tables 2 and 3 show the R^2 values for both predictors in each cohort and for each outcome measure. The curriculum-sampling test score was a moderate-to-strong predictor for 1YGPA and a moderate predictor for 3YGPA, and the introduction course grade was a strong predictor for 1YGPA and 3YGPA. An extensive discussion of the predictive validity of the curriculum-sampling tests and the introduction course grade is provided in Niessen et al. (2016). In addition, as recommended by Mattern and Patterson (2013), plots with separate regression lines for males and females are shown in Figures A1 and A2 in the appendix to aid the interpretation of the results. The plots show that when the regression lines did not overlap, female performance was mostly underpredicted, and more so for lower scores.

Frequentist Step-Down Regression Analyses

Table 2 shows the frequentist results for the step-down regression analyses. Table 3 shows the corresponding d_{Mod} effect sizes. In all analyses, we checked for influential cases by inspecting Cook's distance, but no problematic values were found (all < 1).

Curriculum-sampling test. Statistically significant differential prediction with slope differences and intercept differences was found in the 2013 and 2015 cohorts. However,

Table 2. Frequentist Step-Down Regression Analysis Results

			Full Model						
Criterion	**Predictor**	**Cohort**	***B* IV**	***B* gender**	***B* IV × gender**	R^2 **IV**	ΔR^2 **Omnibus**	ΔR^2 **Slope**	ΔR^2 **Intercept**
1YGPA	Curriculum sample	2013	.12* [.11, .14]	.37* [.018, .57]	–.05* [–.09, –.02]	.26* [.20, .32]	.030* [.001, 064]	.010* [<.001, .034]	.016* [.002, .042]
		2014	.12* [.10, .14]	.08 [–.12, .29]	.01 [–.03, .04]	.23* [.17, .29]	.001 [<.001, .015]		
		2015	.13* [.11, .15]	.23* [.01, .45]	–.05* [–.11, <–.01]	.21* [.14, .27]	.012* [.001, .044]	.007* [<.001, .030]	.008* [<.001, .034]
	Introduction course grade[a]	2013	.68* [.62, .73]	.29* [.14, .44]	–.18* [–.28, –.07]	.56* [.50, .63]	.021* [.006, .047]	.008* [<.001, .025]	.013* [.002, .032]
		2014	.60* [.55, .66]	.08 [–.09, .26]	.09 [–.03, .20]	.48* [.42, .54]	.004 [<.001, .020]		
		2015	.53* [.48, .59]	.11 [–.06, .29]	<.01 [–.12, .12]	.43* [.37, .50]	.002 [<.001, .022]		
3YGPA	Curriculum sample	2013	.06* [.05, .08]	.08 [–.06, .22]	.01 [–.02, .04]	.15* [.09, .21]	.003 [<.001, .026]		
	Introduction course grade	2013	.38* [.33, .43]	.06 [–.06, .18]	–.04 [–.15, .06]	.35* [.35, .42]	.002 [<.001, .021]		

[a]For these analyses, the grade in this course was excluded from the GPA calculations.
*$p < .05$. Men were coded 0, women were coded 1. All IV's were centered around the mean before analyses. All regression coefficients are unstandardized coefficients. 95% confidence intervals are between brackets.

Table 3. d_{Mod} Effect Sizes for Differential Prediction

Criterion	Predictor	Cohort	d_{Mod_signed}	$d_{Mod\_unsigned}$	d_{Mod_under}	d_{Mod_over}	$d_{Mod\_max.}$	$Prop._{\_under}$	$Prop._{\_over}$
1YGPA	Curriculum sample	2013	–.25	.26	–.26	<.01	–.84	.93	.07
		2014	–.06	.06	–.06	<.01	–.12	.98	.02
		2015	–.20	.22	–.21	.01	–.80	.86	.14
	Introduction course[a]	2013	–.18	.19	–.18	<.01	–.61	.92	.08
		2014	–.06	.10	–.08	.02	.35	.72	.28
		2015	–.09	.09	–.09	.00	–.09	1.00	.00
3YGPA	Curriculum sample	2013	–.10	.10	–.10	<.01	–.19	.99	.01
	Introduction course	2013	–.07	.08	–.08	<.01	–.24	.91	.09

Note: The focal group was female, the referent group was male.
[a]For these analyses, the grade in this course was excluded from the GPA calculations. d_{Mod_signed} = the signed effect size for over- and underprediction of female criterion scores based on the male group regression line (can cancel out to 0 when there are slope differences). $d_{Mod\_unsigned}$ = the unsigned effect size for over- and underprediction of female criterion scores based on the male group regression line (does not cancel out when there are slope difference, shows magnitude, but not direction). d_{Mod_under} = the effect size for underprediction in the score range where underprediction for females occurred. d_{Mod_over} = the effect size for overprediction in the score range where underprediction for the females occurred. $d_{Mod\_max.}$ = the largest absolute-value differences between groups' regression lines. $Prop._{\_over}$ = the proportion of overpredicted criterion scores for females. $Prop._{\_under}$ = the proportion of underpredicted criterion scores for females.

the increases in explained variance for slope differences ($\Delta R^2_{2013} = .010$ and $\Delta R^2_{2015} = .007$) and intercept differences ($\Delta R^2_{2013} = .016$ and $\Delta R^2_{2015} = .008$) were small in both samples. For the 2014 sample, the omnibus test did not show statistical evidence in favor of differential prediction. As shown in Table 3, in the cohorts where differential prediction was detected, the signed and unsigned d_{Mod} effect sizes were very similar ($d_{Mod\_signed} = –.25$ and $d_{Mod\_unsigned} = .26$ in 2013, and $d_{Mod\_signed} = –.20$ and $d_{Mod\_unsigned} = .22$ in 2015) which indicated that the effects did not cancel out across the score range due to slope differences. Most female criterion scores (93% and 86%) were underpredicted, with $d_{Mod\_under} = –.26$ in 2013 and $d_{Mod\_under} = –.21$ in 2015. Female overprediction was trivial in both samples, with a maximum effect size of $d_{Mod\_over} = .01$ or smaller. The effect sizes indicated that overall, the detected slope and intercept differences resulted in small female underprediction, while the largest absolute value differences were large ($d_{Mod\_max.} = –.84$ in 2013, and –.80 in 2015). In the 2014 cohort, where no differential prediction was detected, the effect sizes were very small ($d_{Mod\_signed} = –.06$, $d_{Mod\_unsigned} = .06$, $d_{Mod\_max.} = –.12$).

For 3YGPA as the criterion, no statistically significant differential prediction was detected based on the curriculum-sampling test ($\Delta R^2_{omnibus} = .003$). The corresponding effect sizes were small as well ($d_{Mod\_signed} = –.10$, $d_{Mod\_unsigned} = .10$, $d_{Mod\_max.} = –.19$).

Introduction course grade. For the introduction course grade as a predictor of 1YGPA in the 2013 sample, statistically significant differential prediction with slope and intercept differences was found. Again, the increases in explained variance for slope differences ($\Delta R^2 = .008$) and intercept differences ($\Delta R^2 = .013$) were small. The corresponding signed and unsigned d_{Mod} effect sizes were, again, very similar ($d_{Mod\_signed} = –.18$ and $d_{Mod\_unsigned} = .19$), so the effects did not cancel out across the score range due to slope differences. Most female criterion scores (92%) were underpredicted, with $d_{Mod\_under} = –.18$. Overall, female 1YGPA was slightly underpredicted by the introduction course grade, but the largest absolute value difference was of moderate size ($d_{Mod\_max.} = –.61$). For the 2014 and 2015 cohorts, no statistical evidence in favor of differential prediction was found ($\Delta R^2_{omnibus} = .004$, and $\Delta R^2_{omnibus} = .002$, respectively), and the d_{Mod} effect sizes were small as well (2014: $_{dMod\_signed} = –.06$, $d_{Mod\_unsigned} = .10$, $d_{Mod\_max.} = .35$, and 2015: $d_{Mod\_signed} = –.09$, $d_{Mod\_unsigned} = .09$, $d_{Mod\_max.} = –.09$).

Again, for 3YGPA as the criterion, no statistically significant differential prediction was detected based on the introduction course grade ($\Delta R^2_{omnibus} = .002$). The corresponding d_{Mod} effect sizes were small ($d_{Mod\_signed} = –.07$, $d_{Mod\_unsigned} = .08$, $d_{Mod\_max.} = –.24$).

Bayesian Step-Down Regression Analyses

The Bayesian results are shown in Table 4 and they were similar to the frequentist results in most cases. However, the added value of the Bayesian analyses is that, opposed to the frequentist analyses, the results can also show the strength of the evidence in favor of no differential prediction, compared to the strength of the evidence in favor of differential prediction.

Curriculum-sampling test. For the curriculum-sampling test predicting 1YGPA, there was very strong evidence in favor of differential prediction in the 2013 cohort ($BF_{10} = 687.95$) with positive evidence for slope differences ($BF_{10} = 5.74$) and strong evidence for intercept differences ($BF_{10} = 108.61$). The increase in explained variance for slope differences ($\Delta R^2 = .011$, 95% BCI [.001, .030]) and intercept differences ($\Delta R^2 = .018$, 95% BCI [.004, .041]) were, again, small. Thus, corresponding to the frequentist results, there was strong evidence for small intercept and slope differences. For the 2014 sample, the evidence was strongly in favor of no differential prediction ($BF_{10} = .01$). For the 2015 sample, the evidence based on the omnibus test was slightly in favor of no differential prediction ($BF_{10} = .41$, $\Delta R^2_{omnibus} = .021$, 95% BCI [.004, .052]). This differs from the frequentist results, where statistically significant slope differences and intercept differences were detected.

For predicting 3YGPA, the evidence was strongly in favor of no differential prediction based on curriculum-sampling test scores ($BF_{10} = .03$, $\Delta R^2_{omnibus} = .003$, 95% BCI [<.001, .017]).

Introduction course grade. For predicting 1YGPA in the 2013 sample, there was very strong evidence in favor of differential prediction ($BF_{10} = 1487.92$), with positive evidence

Table 4. Bayesian Step-Down Regression Analysis Results for Each Predictor–Criterion Combination in Each Cohort

			Full Model									
Criterion	**Predictor**	**Cohort**	***B* IV**	***B* gender**	***B* IV × gender**	R^2 **IV**	ΔR^2 **Omnibus**	ΔR^2 **Slope**	ΔR^2 **Intercept**	BF_{10} **Omnibus**	BF_{10} **Slope**	BF_{10} **Intercept**
1YGPA	Curriculum sample	2013	.12 [.11, .14]	.37 [.18, .56]	−.05 [−.010, −.02]	.26 [.22, .29]	.015 [.003, .037]	.011 [.001, .030]	.018 [.004, .041]	687.95	5.74	108.61
		2014	.12 [.10, .14]	.08 [−.12, .29]	<.01 [−.03, .04]	.23 [.20, .27]	.003 [<.001, .013]			.01		
		2015	.13 [.11, .15]	.23 [.01, .45]	−.05 [−.11, <−.01]	.21 [.17, .24]	.021 [.004, .052]			.41		
	Introduction course grade[a]	2013	.68 [.62, .73]	.29 [.14, .44]	−.18 [−.28, −.08]	.56 [.52, .62]	.021 [.006, .047]	.008 [.001, .021]	.011 [.003, .025]	1487.92	13.68	64.88
		2014	.60 [.55, .66]	.08 [−.08, .26]	.09 [−.02, .20]	.48 [.43, .52]	.005 [<.001, .09]			.02		
		2015	.53 [.48, .59]	.11 [−.06, .29]	<.01 [−.12, .12]	.43 [.39, .48]	.003 [<.001, .018]			.01		
3YGPA	Curriculum sample	2013	.06 [.05, .08]	.08 [−.06, .22]	<.01 [−.02, .04]	.15 [.12, .19]	.003 [<.001, .017]			.03		
	Introduction course grade	2013	.38 [.33, .43]	.06 [−.06, .18]	−.05 [−.15, .06]	.35 [.30, .39]	.004 [<.001, .018]			.02		

[a]For these analyses, the grade in this course was excluded from the GPA calculations. BF_{10} = Bayes factor for the evidence in favor of the alternative hypothesis relative to the null hypothesis. Men were coded 0, women were coded 1. All IV's were centered around the mean before analyses. All regression coefficients are unstandardized coefficients. 95% credible intervals are between brackets.

for slope differences ($BF_{10} = 13.68$) and strong evidence for intercept differences ($BF_{10} = 64.88$). The increases in explained variance were however, small for slope differences ($\Delta R^2 = .008$, 95% BCI [.001, .021]) and for intercept differences ($\Delta R^2 = .011$, 95% BCI [.003, .025]). For the 2014 and 2015 samples, the evidence was strongly in favor of no differential prediction ($BF_{10} = .02$, and $BF_{10} = .01$, respectively). The frequentist analyses yielded corresponding results.

For predicting 3YGPA, the evidence was strongly in favor of no differential prediction based on the introduction course grades ($BF_{10} = .02$, $\Delta R^2_{omnibus} = .004$, 95% BCI [<.001, .018]).

Discussion

In this study, we investigated differential prediction by gender using a samples approach (Wernimont & Campbell, 1968) to admission testing, based on data obtained in a real admissions context. We expected that differential prediction would be small or nonexistent for curriculum samples. Because a curriculum sample is representative for the criterion, in content, form, and in the preparation that is required, it should tap into KSAOs that are relevant for successful academic performance. The underrepresentation of certain KSAOs is one of the main explanations for differential prediction by gender in admission procedures based on traditional admission tests (Keiser et al., 2016; Kling et al., 2012; Mattern et al., 2017; Stricker et al., 1993). Therefore, we also expected that a more comprehensive "curriculum sample," in the form of the introduction course grade, would show even less differential prediction as compared to the curriculum-sampling admission test, which was designed as a small-scale version of the *introduction to psychology* course.

Taking all results into account, there was evidence in favor of the null hypothesis that differential prediction did not occur in five of the eight predictor–criterion combinations that we studied. We found evidence in favor of differential prediction with slope and intercept differences for the curriculum-sampling test and for the introduction course grade predicting 1YGPA, both in the same cohort (2013). The result was somewhat inconclusive for the curriculum-sampling test predicting 1YGPA in the 2015 cohort. However, in all cases, the effect sizes as indicated by the increases in explained variance were small. As observed by the d_{Mod} effect sizes and the figures in the appendix, the detected slope and intercept differences led to small female underprediction of 1YGPA. The large Bayes factors combined with the small effect sizes may seem contradictory, but this can be interpreted as strong evidence for an effect that has small credible values (e.g., Kruschke & Liddell, 2018). For predicting 3YGPA, the evidence was strongly in favor of no differential prediction for both predictors. In addition, the increases in explained variance for slope and intercept differences and the d_{Mod} effect sizes were mostly somewhat smaller when using the introduction course grade as a predictor, compared to using the curriculum-sampling admission test. This is in line with our expectation that more comprehensive curriculum samples may lead to less differential prediction by gender.

Differential prediction can have several different causes. It is useful to also take score differences into account when interpreting differential prediction findings. We used the guidelines provided by Meade and Fetzer (2009) to interpret our findings, taking score differences into account. In the 2013

cohort, where strong evidence for differential prediction was found for both predictors when 1YGPA was the criterion, small differences in criterion scores (1YGPA) were found between males and females. For the curriculum-sampling tests, there were no differences in scores between males and females. In this case, intercept differences most likely result from criterion bias or omitted variables, which can occur on the predictor and the criterion side. One plausible explanation is that the curriculum-sampling test was not representative enough for the criterion performance. For the introduction course grade as the predictor, the criterion score differences were accompanied by proportional predictor score differences between males and females. Intercept differences with proportional score differences in the dependent and the independent variable are often related to imperfect test or criterion reliability. Slope differences are considered to indicate differential validity (Meade & Fetzer, 2009). However, the effect sizes for the slope differences were consistently very small. Furthermore, no differential prediction was detected for the same predictor scores in the same cohort when 3YGPA was used as the criterion, which also indicated that bias in the predictor scores may not be the main cause of these findings.

Theoretical and Practical Contributions

This was the first study that investigated differential prediction by gender for curriculum-sampling tests. The results were somewhat mixed, with evidence in favor of differential prediction in one cohort, evidence in favor of no differential prediction in another cohort, and a somewhat inconclusive result in another cohort. However, in all cases, the effect sizes were small with smaller effect sizes for more comprehensive curriculum samples. So, using comprehensive representative performance samples that tap into different relevant KSAOs to a larger extent, and require more prolonged effort, may be a method that yields minimal differential prediction. However, very few studies on the "construct saturation" of curriculum samples and other sample-based assessment have been conducted (Lievens & De Soete, 2012). How and to what extent curriculum samples, and other sample-based methods used for prediction in general, tap into relevant KSAOs is a topic that deserves more attention in future research. Such research may provide more insight into the underlying mechanisms of the predictive validity of sample-based assessments and the presence or absence of subgroup differences in scores and in prediction.

As one reviewer noted, another aspect that is important to recognize is the location of the predictor and criterion variables on the "typical performance" versus "maximum performance" continuum (Sackett, Zedeck, & Fogli, 1988). Whereas we would categorize both exam performance and admission test performance as maximum performance measures, admission test performance may represent maximum performance to a larger extent as compared to exam performance. Nevertheless, scores on curriculum-sampling admission tests may resemble "typical" educational performance more closely, as compared to other, more general cognitively oriented admission tests. This may be another potential explanation for their high predictive validity and limited differential prediction.

We also note that, given the different types of analyses and effect sizes currently used in differential prediction studies, it is very difficult to compare results to findings from other studies. Therefore, we encourage other researchers to use the d_{Mod} effect sizes in future differential prediction studies, because these effect sizes can easily be compared and aggregated across studies (Dahlke & Sackett, 2018).

The results presented above are not only of theoretical interest, but also are of practical value. Performance or behavioral sampling, implemented as curriculum samples in educational selection or as work samples in personnel selection, may provide a practically applicable solution to the omitted variables problem. As Niessen and Meijer (2017) discussed, this approach may serve as an alternative to using separate measures for cognitive and noncognitive skills to predict performance. In assessments based on performance sampling, skills and behavioral tendencies are captured through shown behavior and performance, in a relevant context. This provides an advantage over the use of self-report instruments. It may be practically challenging to implement comprehensive and representative curriculum samples that require prolonged effort and investment as admission instruments. However, using distance learning and digital tools such as video lectures or the MOOC format may enable this type of admission instruments (see Reibnegger et al., 2010; Vihavainen et al., 2013).

A second aim of this study was to demonstrate the use of a Bayesian approach to analyze differential prediction. As we discussed and illustrated, the Bayesian approach offered several advantages. Evidence in favor of the null hypotheses could be investigated; intervals of credible values for parameters could be computed. Furthermore, the absence of findings that indicate differential prediction based on frequentist analyses do not warrant the conclusion that there is no differential prediction. Bayesian analysis does allow inspecting evidence in favor of the null hypothesis, relative to an alternative hypothesis. Because statistical analyses are crucial for the interpretation of research results on the basis of which theories are being constructed and practices evaluated, we hope that our analysis and results may inspire other researchers to consider a Bayesian approach in further educational and organizational research.

Limitations

This study was conducted with samples of applicants for a psychology program, so the results should be replicated in other disciplines in future studies. However, this study can serve as a first step to investigate differential prediction in sample-based assessments in education. In addition, we only studied differential prediction by gender. Differential prediction by ethnicity and socioeconomic background are also major challenges in educational measurement (Aguinis et al., 2010; Mattern & Patterson, 2013). However, it is less clear what the basis of differential prediction is based on those variables. Therefore, it is not completely clear if curriculum sampling could be helpful in those situations.

Another possible limitation was that the time that elapsed between measuring the predictor and the criterion was not the same for the two predictors that we studied. The curriculum-sampling test was completed as an admissions exam, while the introduction course grade was obtained after starting the program a few months later. So, for the more comprehensive predictor, less time also elapsed before obtaining the criterion measures, which may affect the results. However, since differential prediction effect sizes were smaller for both predictors when predicting 3YGPA, compared to predicting 1YGPA, the conclusion that differential prediction is smaller

when less time elapsed between measuring the predictor and the criterion is not likely. We also note that we were only able to study differential prediction using 3YGPA as a criterion for one cohort. In addition, there were some missing values in our data, and we applied listwise deletion in those cases. However, because the percentages of missing values were very small, we do not suspect that this affected the results very much.

Adopting a curriculum-sampling approach also has limitations. Curriculum sampling is rather easy to implement in discipline-specific educational selection because the criterion performance is relatively easy to translate into predictor tasks. However, in situations where the criterion behavior is more complex or diverse, such as in vocational education aimed at practical competencies or in colleges where student do not apply to a program in a specific discipline, it may be more challenging or even unfeasible to develop and administer curriculum-sampling instruments for large-scale assessment. These areas deserve more attention in future research. Still, there are many situations in which curriculum sampling can be applied, such as in admission procedures for most European higher education programs, and graduate- and specialized programs in the United States, such as (pre-)medicine and engineering.

Conclusion

Based on our results we tentatively conclude that comprehensive curriculum sampling may offer a practically feasible approach to admission testing that may yield little or no female underprediction, without having to rely on easily fakeable self-report measures, and while maintaining high predictive validity, potentially leading to fairer admission procedures and selection decisions. Advocating the use of performance samples, Asher and Sciarrino (1974) reasoned that the more the predictor and the criterion are alike, the higher the predictive validity will be. Analogously, we suggest that the more the predictor and the criterion are alike, the smaller differential prediction will be.

Appendix

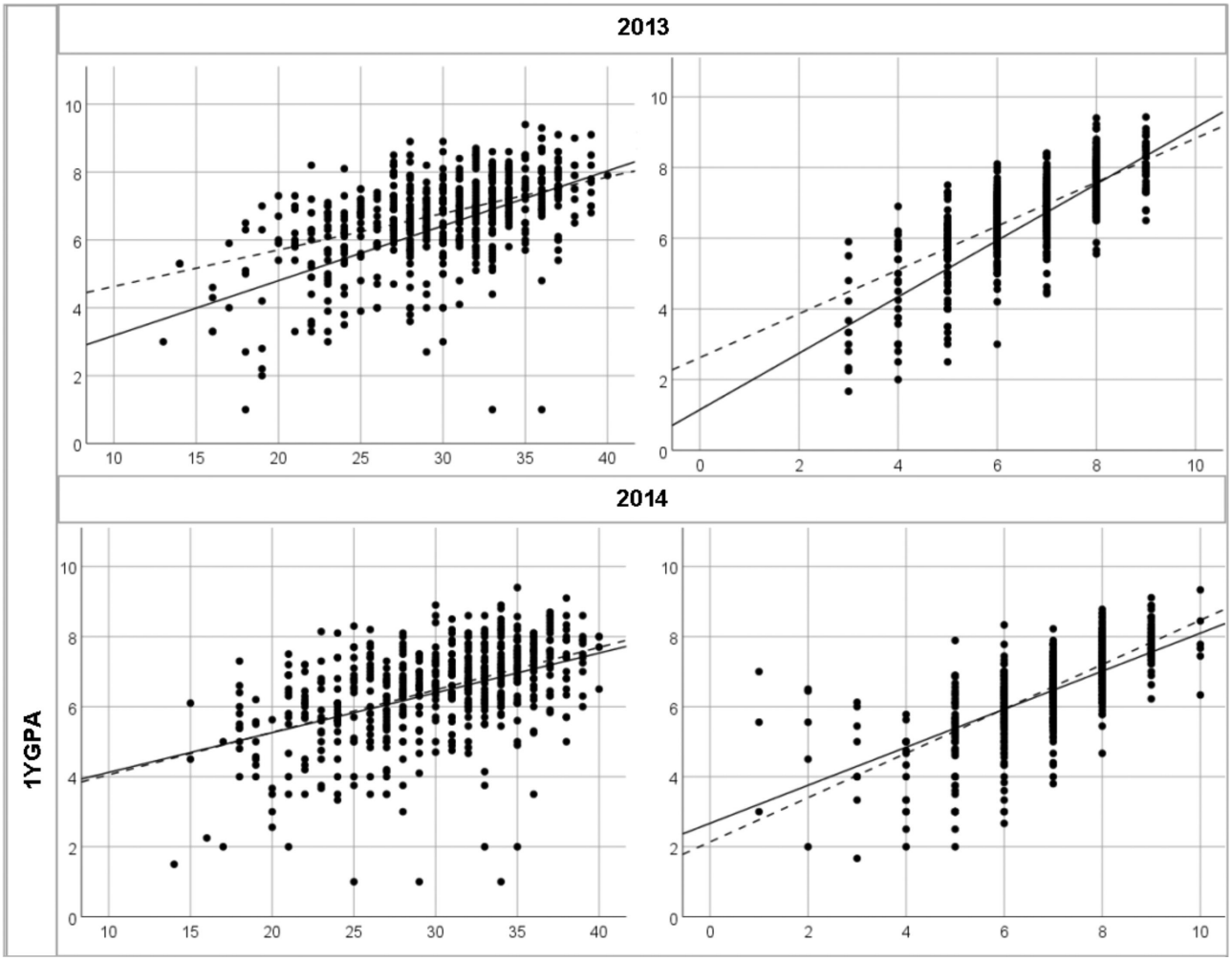

FIGURE A1. Regression plots with separate regression lines for males and females for predicting 1YGPA.

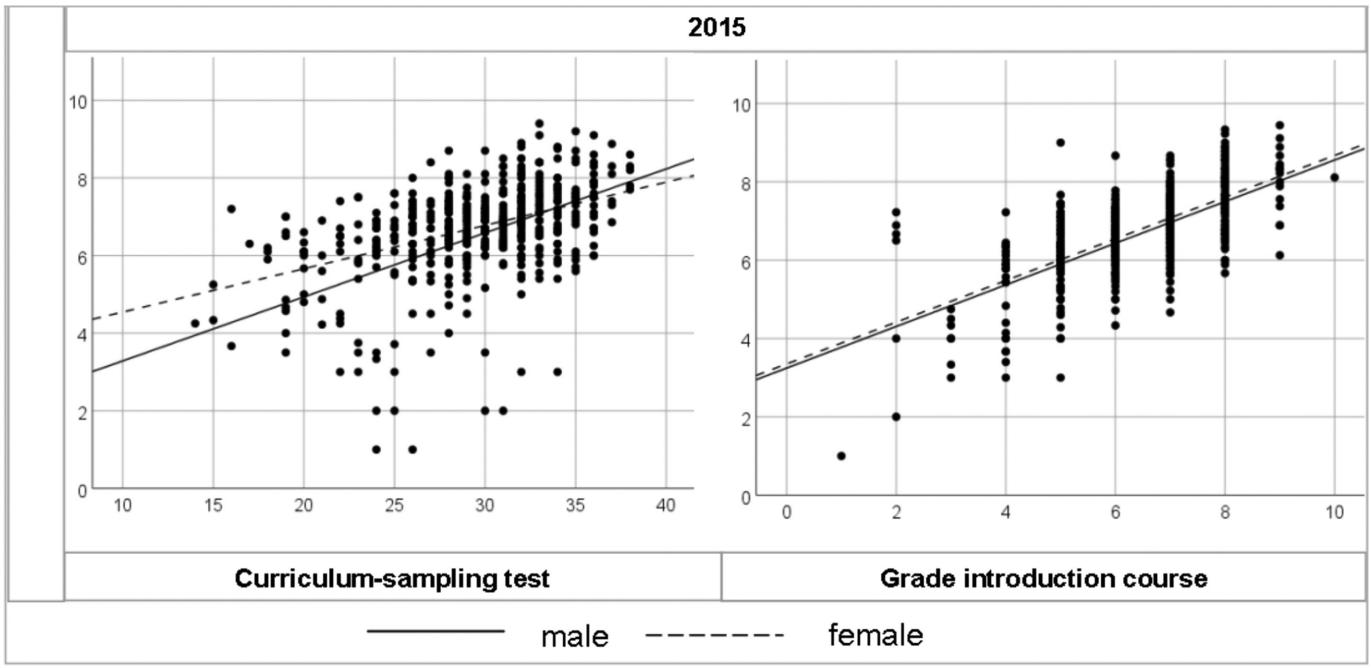

FIGURE A1. *Continued.*

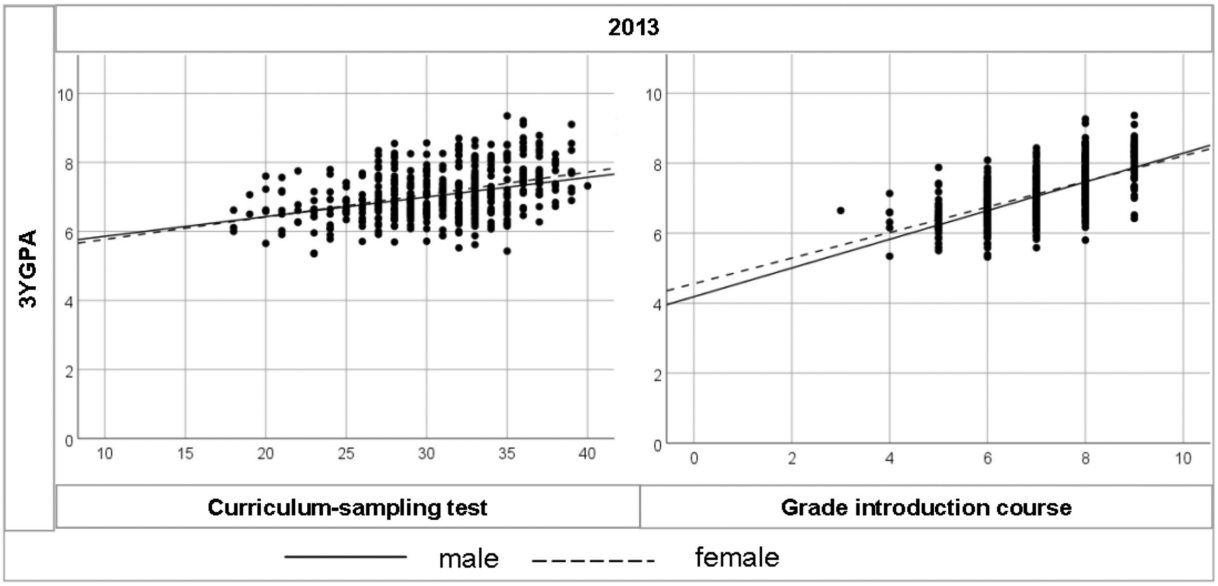

FIGURE A2. Regression plots with separate regression lines for males and females for predicting 3YGPA.

Notes

[1]The term *noncognitive characteristics* often refers to characteristics like personality traits, motivation, and study skills. This term incorrectly implies that these characteristics are completely independent of cognitive skills (e.g., Borghans, Golsteyn, Heckman, & Humphries, 2011; von Stumm & Ackerman, 2013). However, we use this term for simplicity.

[2]In Europe, high school grades are the most common admission criterion (Cremonini, Leisyte, Weyer, & Vossensteyn, 2011), and standardized tests such as the SAT or ACT are not used in most countries. In addition, applicants usually apply to specific academic programs (e.g., medicine, law), instead of to a college.

[3]Applicants could apply to more programs at once, and many withdrew or chose another program, mostly before, and sometimes after they took the admission tests. However, after the application deadline, but before applicants take their admission tests, the media reports on the initial number of applicants and greatly exaggerate the chance to get rejected (e.g., Bouma, 2017).

[4]Although there are more refined ways to handle these missing values, we used this method to be consistent across all analyses. Using more refined methods in combination with the BayesFactor R package is not straightforward.

[5]$BF_{10} = 1$–3: anecdotal evidence for H_1 over H_0, $BF_{10} = 3$–20: positive evidence for H_1 over H_0, $BF_{10} = 20$–150: strong evidence for H_1 over H_0, $BF_{10} \geq 150$: very strong evidence for H_1 over H_0. $BF_{01} = 1 / BF_{10}$, values of BF_{10} smaller than 1 indicate evidence in favor of H_0.

References

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology*, *95*, 648–680.

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology*, *108*, 1045–1059.

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aramburu-Zabala Higuera, L. (2001). Adverse impact in personnel selection: The legal framework and test bias. *European Psychologist*, *6*, 103–111.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology*, *27*, 519–533.

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 435–463.

Booij, A. S., & van Klaveren, C. (2017, June). *Trial lectures or admission talks? How to improve students' choice of major*. Paper presented at the Onderwijs Research Dagen [Education Research Days], Antwerp, Belgium.

Borghans, L., Golsteyn, B. H., Heckman, J., & Humphries, J. E. (2011). Identification problems in personality psychology. *Personality and Individual Differences*, *51*, 315–320.

Borghans, L., Golsteyn, B. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *113*, 13354–13359.

Bouma, K. (2017). Studies met numerus fixus zeer populair: twee keer zo veel aanmeldingen als plekken [Selective programs very popular: Two applicants for each slot]. *De Volkskrant*. Retrieved February 28, 2017, from https://www.volkskrant.nl/binnenland/studies-met-numerus-fixus-zeer-populair-twee-keer-zo-veel-aanmeldingen-als-plekken~a4467897/

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2011). *Crossing the finish line: Completing college at America's public universities*. Princeton, NJ: Princeton University Press.

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences*, *29*, 1057–1068.

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment*, *8*, 248–260.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Credé, M., & Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science*, *3*, 425–453.

Cremonini, L., Leisyte, L., Weyer, E., & Vossensteyn, J. J. (2011). *Selection and matching in higher education: An international comparative study*. Enschede, The Netherlands: Center for Higher Education Policy Studies (CHEPS).

Dahlke, J. A., & Sackett, P. R. (2018). Refinements to effect sizes for tests of categorical moderation and differential prediction. *Organizational Research Methods*, *21*, 226–234.

Dahlke, J. A., & Wiernik, B. M. (2018). psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*. Advance online publication. https://doi.org/10.1177/0146621618795933

De Bolle, M., De Fruyt, F., McCrae, R. R., Löckenhoff, C. E., Costa, P. J., Aguilar Vafaie, M. E., . . . Terracciano, A. (2015). The emergence of sex differences in personality traits in early adolescence: A cross-sectional, cross-cultural study. *Journal of Personality and Social Psychology*, *108*, 171–185. https://doi.org/10.1037/a0038497

de Visser, M., Fluit, C., Fransen, J., Latijnhouwers, M., Cohen-Schotanus, J., & Laan, R. (2017). The effect of curriculum sample selection for medical school. *Advances in Health Sciences Education*, *22*, 43–56.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13–21.

Duckworth, A. L., & Seligman, M. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*, 198–208.

Fischer, F. T., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*, 478–488.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). g: Is this your final answer? *Human Performance*, *15*, 123–142.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.

Häkkinen, I. (2004). *Do university entrance exams predict academic achievement?* (Working Paper No. 2004:16) Department of Economics, Uppsala University. Retrieved November 1, 2017, from https://www.econstor.eu/bitstream/10419/82773/1/wp2004-016.pdf

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*, 152–194.

Jencks, C. (1998). Racial bias in testing. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 55–85). Washington, DC: Brookings Institution Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology*, *101*, 569–581.

Kling, K. C., Noftle, E. E., & Robins, R. W. (2012). Why do standardized tests underpredict women's academic performance? The role of conscientiousness. *Social Psychological and Personality Science*, *4*, 600–606.

Korthals, A. H. (2007). *Commissie 'Ruim baan voor talent'. Eindrapportage [Committee "Room for talent" final report]*. The Hague, The Netherlands: Ministry of Education, Culture, and Science. Retrieved November 1, 2017, from https://www.rijksoverheid.nl/actueel/nieuws/2007/12/11/eindrapportcommissie-ruim-baan-voor-talent

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*, 722–752.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*, 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*, *19*, 339–345.

Lautenschlager, G. J., & Mendoza, J. L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement*, *10*, 133–139.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus & Giroux.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, *10*, 245–257.

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and se-*

lection (pp. 383–410). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199732579.013.0017

Mattern, K. D., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology*, *98*, 134–147.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT* (Research Report No. 2008-4). Princeton, NJ: Educational Testing Service. Retrieved May 18, 2018, from https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2008-4-differential-validity-prediction-sat.pdf

Mattern, K. D., Sanchez, E., & Ndum, E. (2017). Why do achievement measures underpredict female academic performance? *Educational Measurement: Issues and Practice*, *36*(1), 47–57.

Meade, A. M., & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods*, *12*, 738–761.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-2. Retrieved December 22, 2016, from http://CRAN.R-project.org/package=BayesFactor

Niessen, A. S. M., Meijer, R. R. (2017). On the use of broadened admission criteria in higher education. *Perspectives on Psychological Science*. *12*, 436–448.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Predicting success in higher education using proximal predictors. *Plos ONE*, *11*(4): e0153663.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017a). Applying organizational justice theory to admission into higher education: Admission from a student perspective. *International Journal of Selection and Assessment*, *25*, 70–82.

Niessen, A. S. M. & Meijer, R. R., & Tendeiro, J. N. (2017b) Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences*, *106*, 183–189.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLoS ONE*, *13*(6): e0198746.

Nye, C. D., & Sackett, P. R. (2017). New effect sizes for tests of categorical moderation and differential prediction. *Organizational Research Methods*, *20*, 639–664.

Peterson, M. H., Griffith, R. L., Isaacson, J. A., O'Connell, M. S., & Mangos, P. M. (2011). Applicant faking, social desirability, and the prediction of counterproductive work behaviors. *Human Performance*, *24*, 270–290.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racio-ethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*, 153–172.

Plummer, M. (2016a). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling [Computer software].

Plummer, M. (2016b). rjags: Bayesian Graphical Models using MCMC. R package version 4–6. Retrieved December 22, 2016, from http://CRAN.R-project.org/package=rjags

Reibnegger, G., Caluba, H. C., Ithaler, D., Manhal, S., Neges, H. M., & Smolle, J. (2010). Progress of medical students after open admission or admission based on knowledge tests. *Medical Education*, *44*, 205–214.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*, 353–387.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. *Journal of Occupational Psychology*, *55*, 171–183.

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology*, *88*, 1046–1056.

Sackett, P. R., Walmsley, P. T., Koch, A. J., Beatty, A. S., & Kuncel, N. R. (2016). Predictor content matters for knowledge testing: Evidence supporting content validation. *Human Performance*, *29*, 54–71.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*, 482–486.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*, 168–182.

Schult, J., Hell, B., Päßler, K., & Schuler, H. (2013). Sex-specific differential prediction of academic achievement by German ability tests. *International Journal of Selection and Assessment*, *21*, 130–134.

Shewach, O. R., Shen, W., Sackett, P. R., & Kuncel, N. R. (2017). Differential prediction in the use of the SAT and high school grades in predicting college performance: Joint effects of race and language. *Educational Measurement: Issues And Practice*, *36*(3), 46–57.

Strenta, A. C., Elliot, R., Adair, R., Matier, M., & Scott, J. (1994). Choosing and leaving science in highly selective institutions. *Research in Higher Education*, *35*, 513–547.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from scholastic aptitude test scores. *Journal of Educational Psychology*, *85*, 710–718.

Valli, R. & Johnson, P. (2013). Entrance examinations as gatekeepers. *Scandinavian Journal of Educational Research*, *51*, 493–510.

Vihavainen, A., Luukkainen, M., & Kurhila, J. (2013, October). *MOOC as semester-long entrance exam*. Paper presented at the 14th Annual ACM SIGITE Conference on Information Technology Education, Orlando, FL.

Visser, K., van der Maas, H., Engels-Freeke, M., & Vorst, H. (2012). Het effect op studiesucces van decentrale selectie middels proefstuderen aan de poort [The effect on study success of student selection through trial-studying]. *Tijdschrift voor Hoger Onderwijs*, *30*, 161–173.

von Stumm, S., & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin*, *139*, 841–869.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review*, *14*, 779–804.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372–376.

Zwick, R. (2017). *Who gets in? Strategies for fair and effective college admissions*. Cambridge, MA: Harvard University Press

Zwick, R., & Himmelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement*, *48*, 101–121.